sorcero®

The **Ingestum**™ document ingestion framework

Martín Abente Lahaye, Juan Pablo Ugarte, Walter Bender

LibrePlanet, March 21, 2021

What is ingestion?

How can we process the text from unstructured content?

...in a way that is methodical, reusable, extensible, and scalable?

PDFs...

LibreOffice documents...

XML files and web pages..

audio recordings of Jitsi meetings...

email...

Twitter or RSS feeds…

Pubmed or Proquest searches...

scanned documents…

Why is this a challenge?

Text extraction is a fragmented market.

“Digitizing the last 20 percent of documents is particularly challenging.”

There are numerous stand-alone (proprietary) products that address specific extraction scenarios and some broader solutions that claim to offer a general solution to unstructured document ingestion.

How does FOSS help?

There is a plethora of powerful FOSS tools...

… but no extensible framework in which to deploy them.

Beautiful Soup,

Camelot,

PDFMiner,

Pyexcel,

Twython,

Python-tesseract,

Deep Speech,

et al

What is the **Ingestum** approach?

Ingestum

(pronounced “ingest’em”) goes from specific unstructured content sources to general structured outputs.

Ingestum is designed to:

- facilitate writing scripts to extract unstructured content from arbitrary sources;
- provide a framework for extracting content from the diverse universe of sources; and
- allow for the integration with Python scripts and services at many levels of granularity.

How does **Ingestum** work?

Ingestum

components and stages

Sources: source files or data streams, converted to Ingestum Documents for further processing;

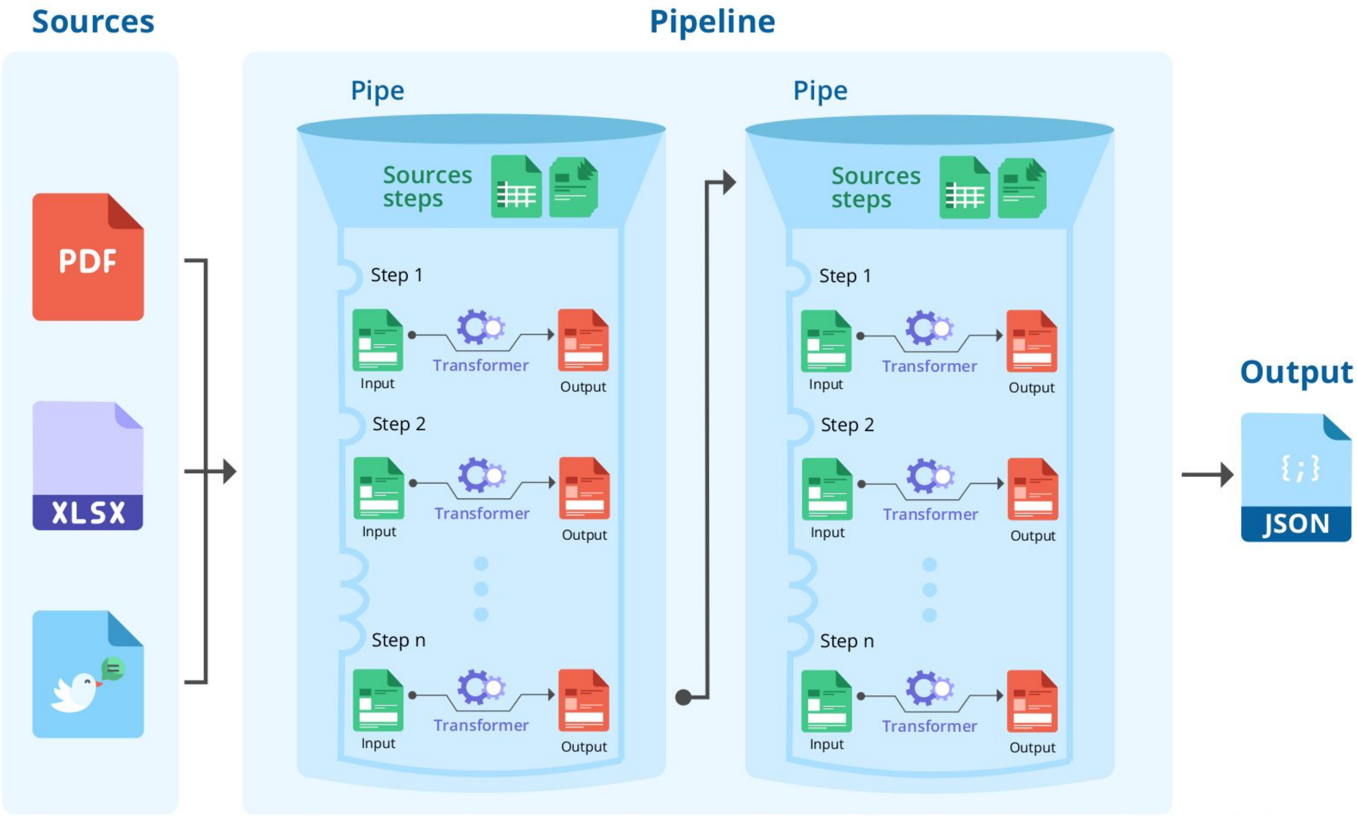
Documents: extensible JSON-encoded formats to which transformations are applied;

Transformers: specific operations on all or part of its input, returning an output Document;

Pipes & Pipelines: a Pipe is a sequence of Transformers, a Pipeline is a collection of Pipes;

Conditionals: logical conditions to apply a Transformer selectively;

Manifests: descriptions of Sources and Pipelines and their parameters.



Sources are fed to a *Pipeline*, which consists of a series of *Pipes*, each of which contain a sequence of *Transformer* steps.

Demonstrations of **Ingestum** in practice

Discussion

```
git clone https://gitlab.com/sorcero/community/ingestum.git
```

https://sorcero.gitlab.io/community/ingestum/

Contact us at:
(martin, xjuan, walter)
@sorcero.com

– fin –